Normalized Ordinal Distance; A Performance Metric for Ordinal, Probabilistic-ordinal or Partial-ordinal Classification Problems

Mohammad Hasan Bahari*, Hugo Van hamme

Center for processing speech and images, KU Leuven, Belgium

Abstract

In this paper, a novel application-independent performance metric for ordinal, probabilistic-ordinal and partial-ordinal classification problems is introduced. Conventional performance metrics for ordinal classification problems, such as mean absolute error of consecutive integer labels and ranked probability score, are difficult to interpret and may lead to fraudulent results about the true performance of the classifier. In this paper, first, the ordinal distance between two arbitrary vectors in Euclidean space is introduced. Then, a new performance metric, namely normalized ordinal distance, is proposed based on the introduced ordinal distance. This performance metric is conceptually simple, computationally inexpensive and application-independent. The advantages of the proposed method over the conventional approaches and its different characteristics are shown using several numerical examples.

Keywords: performance metric, ordinal distance, ordinal classification, probabilistic-ordinal classification, partial-ordinal classification

1. INTRODUCTION

Classification problems can be roughly divided into nominal and ordinal. In nominal classification, the category labels are name-based and have no ranking relation with each other. For example in language recognition problem the category labels are the language names (Bahari et al., 2014). In ordinal classification, there is an intrinsic ordering between the categories. For example, in quality prediction systems, the task is to categorize the quality of a product into bad, good and excellent (Erdural, 2006). In human age group recognition from speech or images, the categories can be child, young, middle-aged and senior (Bahari and Van hamme, 2011b; Li et al., 2013). In the classification of the therapeutic success, the classes are good recovery, moderate disability, severe disability, and fatal outcome (Cardoso and da Costa, 2007). In all ordinal classification problems (C_O), the class labels are ordinal numbers, i.e. there is intrinsic ordering between the categories.

Probabilistic-Ordinal and Partial-Ordinal Classification problems, labeled as C_O^{Pr} and C_O^{Pa} respectively, are well-known generalizations of the C_O. In C_O^{Pr}, for a test

*Corresponding author. Tel:+32-(0)16-32.85.45. Fax:+32-(0)16-32.17.23.

Email addresses: mohamadhasan.bahari@esat.kuleuven.be (Mohammad Hasan Bahari), hugo.vanhamme@esat.kuleuven.be (Hugo Van hamme)

datapoint, the classifier calculates the probability of belonging to each category. In C_O^{Pa}, instead of the crisp class labels each datapoint has a degree of membership to every class (Verwaeren et al., 2012). These types of problems, explained in sections 2.2 and 2.3 in detail, can be found in many domains, such as natural language processing, social network analysis, bioinformatics and agriculture (Verwaeren et al., 2012).

Scientists have proposed different methods to solve C_O, C_O^{Pr} and C_O^{Pa} (Verwaeren et al., 2012; McCullagh, 1980; Chu and Keerthi, 2007; Cheng et al., 2008; Chu and Ghahramani, 2004; Shevade and Chu, 2006). For example, McCullagh introduced an ordinal classifier, namely the proportional odds model (POM), based on logistic regression (1980). In (Chu and Keerthi, 2007), C_O is addressed using a generalization of support vector machines (SVM) namely support vector ordinal regression (SVOR). A neural network approach for the C_O is suggested in (Cheng et al., 2008). Chu and Ghahramani (2004) suggested Gaussian processes for C_O. In (Verwaeren et al., 2012), kernel-based proportional odds models is introduced to solve the C_O^{Pa}.

To measure the performance of these classifiers, different approaches have been suggested. For example, mean zero-one error (E_{mzo}) and mean absolute error of consecutive integer labels (E_{ma}^{cil}) are widely applied to measure the performance of the classifiers in C_O (Chu and Keerthi, 2007; Cheng et al., 2008; Chu and Ghahramani, 2004; Shevade and Chu, 2006). However, non of these methods are applicable to C_O^{Pr} and C_O^{Pa}. Percentage of correctly fuzzy classified instances (P_{cfci}) and Average Deviation (E_{ad}) have been suggested to measure the classifier performance in C_O^{Pr} and C_O^{Pa} (Verwaeren et al., 2012; Manel et al., 2002; Van Broekhoven et al., 2007; Mouton et al., 2009). The main drawback of P_{cfci} is that it does not consider the order of categories (Manel et al., 2002; Van Broekhoven et al., 2007). The E_{ad} suggests a simple idea to solve this problem (Van Broekhoven et al., 2007; Mouton et al., 2009). Although the E_{ad} is attractive from several aspects, the interpretation of its results is difficult, because the range of its output depends on the application. The same difficulty is observed in E_{ma}^{cil}. Application dependency makes the interpretation of E_{ma}^{cil} and E_{ad} very challenging. The average of ranked probability scores (E_{rps}), is also applied as a performance metric in C_O^{Pr} and C_O^{Pa} (Bougeault, 2003; Murphy, 1969). In this method, the order of categories is important and the range of the output is fixed between 0 and 1. This method can be applied to C_O, C_O^{Pr} and C_O^{Pa}. However, analysis reveals that E_{rps} over estimates the performance of classifiers in many situations. This issue, which leads to a erroneous interpretation of classifier performance, is illustrated by some numerical examples in section 5.

In this paper, we investigate different characteristics of these performance metrics and finally a novel application-independent performance metric, namely Normalized Ordinal Distance (E_{nod}^{p}), is introduced. The Matlab code of the suggested approach, which can be applied to all three types of considered problems C_O, C_O^{Pr} and C_O^{Pa}, can be downloaded from our website[1].

This paper is organized as follows. In section 2, the mathematical formulations of C_O, C_O^{Pr} and C_O^{Pa} are presented. In section 3, five different conventional performance metrics are explained. The proposed performance metric is elaborated in section 4. In section 5, the effectiveness of the proposed approach is illustrated using some numerical examples. The paper endes with a conclusion in section 6.

[1] http://www.esat.kuleuven.be/psi/spraak/downloads/

2. PROBLEM FORMULATION

In this section, the ordinal, probabilistic-ordinal and partial-ordinal problems are formulated.

2.1. ORDINAL CLASSIFICATION

Assume that we are given a training data set $S^{\text{tr}} = \{(X_1, Y_1), \cdots, (X_n, Y_n), \cdots, (X_N, Y_N)\}$, where $X_n = [x_{n,1}, \cdots, x_{n,i}, \cdots, x_{n,I}]$ denotes a vector of observed characteristics of the data item and $Y_n = \left[y_{n,1}, \cdots, y_{n,d}, \cdots, y_{n,D}\right]$ denotes a label vector. The label vector is defined as follows if X_n belongs to class C_d.

$$y_{n,j} = \begin{cases} 1 & j = d \\ 0 & j \neq d \end{cases} \tag{1}$$

In ordinal problems, there is an intrinsic ordering between the classes, which is denoted as $C_1 \prec \cdots \prec C_d \prec \cdots \prec C_D$ like low, medium and high (Verwaeren et al., 2012). The goal is to approximate a classifier function (G), such that for the m^{th} unseen observation X_m^{tst}, $\hat{Y}_m = G(X_m^{tst})$ is as close as possible to the true label. For a crisp classifier $\hat{Y}_m$ is defined as follows if the d^{th} class is chosen for X_m^{tst}.

$$\hat{y}_{m,j} = \begin{cases} 1 & j = d \\ 0 & j \neq d \end{cases} \tag{2}$$

2.2. PROBABILISTIC-ORDINAL CLASSIFICATION

Probabilistic-ordinal classification problem ($\text{C}_\text{O}^{\text{Pr}}$) is a generalization of the C_O, where each element of the classifier output vector ($\hat{Y}$) represents the probability of belonging to the corresponding category. In this type of classification, Y_n is defined by relation (1). However, $\hat{Y}_m$ is defined as follows.

$$\hat{Y}_m = \left\{ \left[\hat{y}_{m,1}, \cdots, \hat{y}_{m,d}, \cdots, \hat{y}_{m,D}\right] \in \mathbb{R}^D \middle| \hat{y}_{m,d} \geq 0; \textstyle\sum_{d=1}^{D} \hat{y}_{m,d} = 1 \right\} \tag{3}$$

where $\mathbb{R}$ denotes the set of real numbers.

2.3. PARTIAL-ORDINAL CLASSIFICATION

Partial-ordinal classification problem ($\text{C}_\text{O}^{\text{Pa}}$) is another generalization of C_O (Verwaeren et al., 2012). In ordinal problems, each data object is limited to belong to a single category, i.e. out of all D elements of Y_n , only one is nonzero. However, this is too conservative in the case of non-crisp or fuzzy classes. This limitation is relaxed in $\text{C}_\text{O}^{\text{Pa}}$ by rephrasing Y_n as follows.

$$Y_n = \left\{ \left[y_{n,1}, \cdots, y_{n,d}, \cdots, y_{n,D}\right] \in \mathbb{R}^D \middle| y_{n,d} \geq 0; \textstyle\sum_{d=1}^{D} y_{n,d} = 1 \right\} \tag{4}$$

Therefore, each datapoint has a degree of membership to all classes. Like in ordinal problems, the final goal is to approximate a classifier function (G), such that for an unseen observation X^{tst}, $\hat{Y}_m = G(X_m^{tst})$ is as close as possible to the true label. In this type of classification $\hat{Y}_m$ is also defined by relation 3.

3. CONVENTIONAL PERFORMANCE METRICS

In this section, five widely-used conventional metrics, namely E_{mzo}, $E_{\text{ma}}^{\text{cil}}$, P_{cfci}, E_{ad} and E_{rps} are introduced (Verwaeren et al., 2012; McCullagh, 1980; Chu and Keerthi, 2007; Cheng et al., 2008; Chu and Ghahramani, 2004; Shevade and Chu, 2006; Manel et al., 2002; Van Broekhoven et al., 2007; Mouton et al., 2009; Murphy, 1969; Kohonen and Suomela, 2005; Toda, 1963).

3.1. MEAN ZERO-ONE ERROR (E_{MZO})

Performance metric E_{mzo} is the fraction of incorrect predictions, which is calculated as follows (Chu and Keerthi, 2007; Cheng et al., 2008; Chu and Ghahramani, 2004; Shevade and Chu, 2006).

$$E_{\text{mzo}} = \frac{1}{M}\sum_{m=1}^{M} 1_{\hat{y}_m \neq y_m} \tag{5}$$

where M is the total number of test set datapoints, $\hat{y}_m$ is the predicted label of the m^{th} test set datapoint and y_m is the true label of the m^{th} test set datapoint. The main advantage of E_{mzo} is its simplicity. However, it does not consider the order of the categories. Furthermore, it is not applicable to measure the performance in $\text{C}_{\text{O}}^{\text{Pr}}$ or $\text{C}_{\text{O}}^{\text{Pa}}$.

3.2. MEAN ABSOLUTE ERROR OF CONSECUTIVE INTEGER LABELS ($E_{\text{MA}}^{\text{CIL}}$)

To calculate the $E_{\text{ma}}^{\text{cil}}$, first, both true labels and predicted labels of the test set datapoints are transformed into consecutive integers so that if the d^{th} column of the label vector is 1 then the transformed label is equal to d (Chu and Keerthi, 2007; Cheng et al., 2008; Chu and Ghahramani, 2004; Shevade and Chu, 2006). After label transformation the $E_{\text{ma}}^{\text{cil}}$ is calculated as follows.

$$E_{\text{ma}}^{\text{cil}} = \frac{1}{M}\sum_{m=1}^{M} |\hat{U}_m - U_m| \tag{6}$$

where $\hat{U}_m$ is the transformed predicted label of the m^{th} test set datapoint and U_m is the transformed true label of the m^{th} test set datapoint. The $E_{\text{ma}}^{\text{cil}}$ enjoys the advantage of considering the order of categories into account. However, it cannot be applied to evaluate the classifiers in $\text{C}_{\text{O}}^{\text{Pr}}$ or $\text{C}_{\text{O}}^{\text{Pa}}$ Moreover, the range of its output is application-dependent. Therefore, the interpretation of this metric is challenging. This is shown in section 5 using some numerical examples.

3.3. PERCENTAGE OF CORRECTLY FUZZY CLASSIFIED INSTANCES (P_{CFCI})

Performance metric P_{cfci} has been applied to measure the performance of probabilistic or fuzzy classifiers (Manel et al., 2002; Van Broekhoven et al., 2007). It is calculated as follows:

$$P_{\text{cfci}} = \frac{100}{M}\sum_{m=1}^{M}\left(1 - \frac{1}{2}\sum_{d=1}^{D}|\hat{y}_{m,d} - y_{m,d}|\right) \tag{7}$$

As it can be inferred from the above relation, the order of the categories is not considered in P_{cfci}.

3.4. AVERAGE DEVIATION (E_{AD})

Performance metric E_{ad} was originally introduced by Van Broekhoven (Van Broekhoven et al., 2007) to evaluate the classifiers in fuzzy ordered classification problems. It was also applied in different applications with other names (Verwaeren et al., 2012; Mouton et al., 2009). The E_{ad} is calculated as follows:

$$E_{\mathrm{ad}} = \frac{1}{M}\sum_{m=1}^{M}\left\{\sum_{d=1}^{D-1}\left|\sum_{i=1}^{d}\hat{y}_{m,i} - \sum_{i=1}^{d} y_{m,i}\right|\right\} \tag{8}$$

It can be interpreted from the above relation that the order of categories is important in E_{ad}. E_{ad} is also useful for classifier evaluation in $\mathrm{C_O^{Pr}}$ or $\mathrm{C_O^{Pa}}$. However, similar to $E_{\mathrm{ma}}^{\mathrm{cil}}$, the range of E_{ad} is application-dependent and hence difficult to interpret.

3.5. AVERAGE RANKED PROBABILITY SCORES (E_{RPS})

The ranked probability score was originally introduced to score the output of probabilistic classifiers (Bougeault, 2003; Murphy, 1969). It is defined as follows.

$$RPS_Y(\hat{Y}) = \frac{1}{D-1}\left\{\sum_{d=1}^{D-1}\left(\sum_{i=1}^{d}\hat{y}_i - \sum_{i=1}^{d} y_i\right)^2\right\} \tag{9}$$

This scoring rule can be easily extended to measure the performance of classifiers in $\mathrm{C_O}$, $\mathrm{C_O^{Pr}}$ and $\mathrm{C_O^{Pa}}$ using the following relation.

$$E_{\mathrm{rps}} = \frac{1}{M(D-1)}\sum_{m=1}^{M}\sum_{d=1}^{D-1}\left(\sum_{i=1}^{d}\hat{y}_{m,i} - \sum_{i=1}^{d} y_{m,i}\right)^2 \tag{10}$$

As it can be interpreted from the above relation, the order and the number of categories are important in E_{rps}. It is assumed that the maximum of the nominator of E_{rps} is $M(D-1)$. Therefore, to fix the range of E_{rps} between 0 and 1 the nominator is divided to its maximum possible value $M(D-1)$. However, this assumption is very conservative so that in many practical cases the maximum of the nominator of E_{rps} is less than $M(D-1)$. Consequently, this assumption may lead to an erroneous interpretation of the classifier performance. Numerical examples of Section 5 reveal this issue clearly.

4. PROPOSED PERFORMANCE METRIC

In this section, first, Ordinal Distance (OD) of two vectors in Euclidean space is introduced. Then, a new performance metric, namely normalized ordinal distance (E_{nod}^{p}), is developed based on the ordinal distance.

4.1. ORDINAL DISTANCE (OD)

In this section, the definition of a distance function is recaptured. Then, the Minkowski distance is described and finally, the ordinal distance is introduced as an extension of the Minkowski distance.

4.1.1. Distance

By definition, a distance function of two points $A = [a_1, \cdots, a_d, \cdots, a_D]$ and $B = [b_1, \cdots, b_d, \cdots, b_D]$ is a function $D : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, which satisfies the following three conditions (Deza and Deza, 2009):
1. $D(A,B) \geq 0$ and $D(A,B) = 0 \Leftrightarrow A = B$
2. $D(A,B) = D(B,A)$
3. $D(A,C) \leqslant D(A,B) + D(B,C)$
A variety of distance functions have been introduced by scientists for different applications such as Minkowski distance, Mahalanobis distance, Chebyshev distance and Hamming distance (Deza and Deza, 2009).

4.1.2. The Minkowski Distance of Order p

The Minkowski distance of order p or p-norm is a distance function, which satisfies all conditions of a distance function.

$$||A-B||_p = \left(\sum_{d=1}^{D} |a_d - b_d|^p \right)^{1/p} \tag{11}$$

where p is a real number not less than 1. As in can be interpreted from relation (11), in p-norm, the order of the elements of two points A and B, is not important.

4.1.3. The Ordinal Distance of Order p

The notion of ordinal distance is previously used to measure the differences of two strings (Morovic et al., 2002) or two histograms (Luxenburger, 2008). In this paper, an ordinal distance of two vectors in Euclidean space is introduced. The Ordinal Distance of order p between two points A and B is defined in relation 12.

$$\begin{aligned} ||A-B||_p^{\mathrm{OD}} &= \left(\sum_{d=1}^{D} \left| \bar{a}_d - \bar{b}_d \right|^p \right)^{1/p} \\ \bar{a}_d &= \sum_{i=1}^{d} a_i \\ \bar{b}_d &= \sum_{i=1}^{d} b_i \end{aligned} \tag{12}$$

where p is a real number not less than 1. Since (12) is a Minkowski distance between $\bar{A} = [\bar{a}_1 \cdots \bar{a}_d \cdots \bar{a}_D]$ and $\bar{B} = [\bar{b}_1 \cdots \bar{b}_d \cdots \bar{b}_D]$, it follows that the ordinal distance of order p satisfies the conditions of section 4.1.1.

Figure 1 shows the diagram of unit circle using Minkowski and Ordinal distances of orders 1, 2 and infinity.

4.2. NORMALIZED ORDINAL DISTANCE (E_{NOD}^P)

In this section, a new performance metric, namely normalized ordinal distance (E_{nod}^p), is introduced to measure the performance classifiers in $\mathrm{C_O}$, $\mathrm{C_O^{Pr}}$ and $\mathrm{C_O^{Pa}}$.

$$E_{\mathrm{nod}}^p = \frac{\sum_{m=1}^{M} \left|\left| Y_m - \hat{Y}_m \right|\right|_p^{\mathrm{OD}}}{\sum_{m=1}^{M} \psi_{Y_m}^p} \tag{13}$$

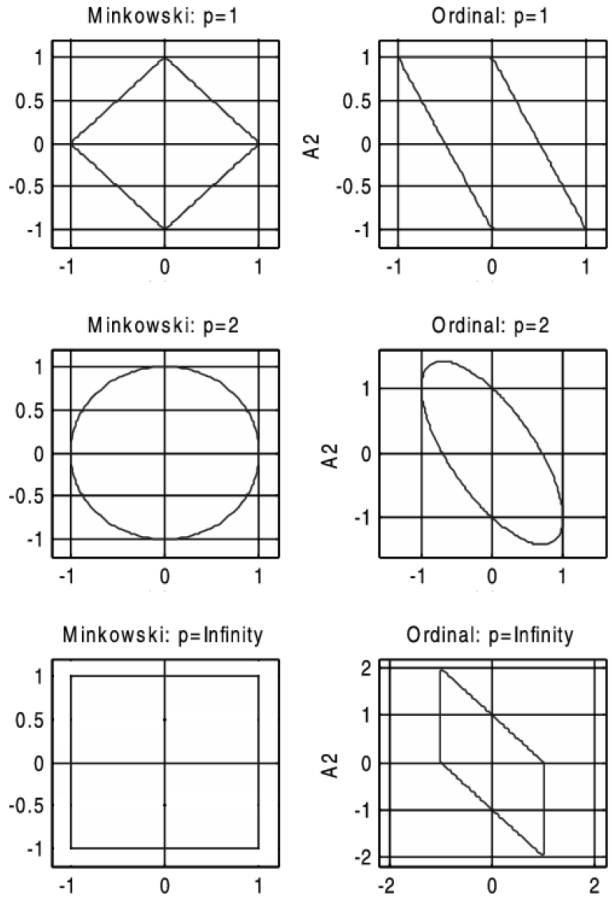

Figure 1: *Diagram of unit circle using Minkowski and Ordinal distances of orders 1, 2 and infinity.*

where $\psi_{Y_m}^p$ is the upper bound of $\|Y-\hat{Y}\|_p^{\mathrm{OD}}$ for any possible $\hat{Y}$ in its defined range. ψ_Y is defined as follows.

$$\psi_Y^p \triangleq \max_T \|Y-T\|_p^{\mathrm{OD}} \tag{14}$$

where $T=\{t_1,\cdots,t_d,\cdots,t_D\}$ is an arbitrary vector with the same specifications of $\hat{Y}$ mentioned in relation (2). ψ_Y^p can be calculated using theorem 1.

In E_{nod}^p ordinal distance is used to take the order of categories into account and it is normalized by the largest possible ordinal distance because not all test cases (Y_m) are equally difficult and the possible ordinal distance for some test cases is larger than others. Without this normalization the ordinal distance is difficult to interpret. In this paper, we are performing a macro-averaging, while a micro-averaging variant could also be studied.

Theorem 1:
The upper bound of $\|Y-\hat{Y}\|_p^{\mathrm{OD}}$ for any possible $\hat{Y}$ can be obtained as follows.

$$\psi_Y^p = max\left(\|Y-L_1\|_p^{\mathrm{OD}},\cdots,\|Y-L_d\|_p^{\mathrm{OD}},\cdots,\|Y-L_D\|_p^{\mathrm{OD}}\right) \tag{15}$$

or equivalently

$$\psi_Y^p = max\left(\|Y-L_1\|_p^{\mathrm{OD}}, \|Y-L_D\|_p^{\mathrm{OD}}\right) \tag{16}$$

where L_d is a vector of size Y. The d^{th} element of L_d is equal to 1 and the rest of elements are zero. As it can be interpreted from relations (15) and (16), although the latter one is more restrictive, it provides an easier way to calculate ψ_Y^p.

Proof:

We first prove the relation (15), which help us to show the correctness of relation (16).

Proof of relation (15):

By definition

$$\|Y-T\|_p^{\mathrm{OD}} = \|\Lambda(Y-T)\|_p \tag{17}$$

where Λ is a lower triangular matrix of size $D \times D$ with all diagonal and lower diagonal elements equal to 1. Since $\|(Y-T)\|_p$ is a convex function of T and a convex function remains convex under an affine transformation, $\|\Lambda(Y-T)\|_p$ is also convex.

On the other hand, a convex function on a compact convex set attains its maximum at an extreme point of the set (Kincaid and Cheney, 2002). In this problem $T \in \{[t_1,\cdots,t_d,\cdots,t_D] \in \mathbb{R}^D | t_d \geq 0; \sum_{d=1}^D t_d = 1\}$. The extreme points of this compact convex set are L_d with $d \in \{1,\cdots,D\}$.

Therefore

$$\max_T \|\Lambda(Y-T)\|_p = max\left(\|\Lambda(Y-L_1)\|_p, \cdots, \|\Lambda(Y-L_d)\|_p, \cdots, \|\Lambda(Y-L_D)\|_p\right) \tag{18}$$

Consequently

$$\max_T \|Y-T\|_p^{\mathrm{OD}} = max\left(\|Y-L_1\|_p^{\mathrm{OD}}, \cdots, \|Y-L_d\|_p^{\mathrm{OD}}, \cdots, \|Y-L_D\|_p^{\mathrm{OD}}\right) \tag{19}$$

Proof of relation (16):

Relation (16) is now shown by contradiction. Suppose relation (15) is not equivalent with relation (16), then there must be a $k \in \{2,\cdots,D-1\}$ such that

$$\|Y-L_k\|_p^{\mathrm{OD}} > \|Y-L_1\|_p^{\mathrm{OD}} \tag{20}$$

$$\|Y-L_k\|_p^{\mathrm{OD}} > \|Y-L_D\|_p^{\mathrm{OD}} \tag{21}$$

Expansion of relation (20) and (21) is

$$\sum_{d=1}^{k-1}(\sum_{i=1}^{d} y_i)^p + \sum_{d=k}^{D-1}(1-\sum_{i=1}^{d} y_i)^p > \sum_{d=1}^{D-1}(1-\sum_{i=1}^{d} y_i)^p \tag{22}$$

$$\sum_{d=1}^{k-1}(\sum_{i=1}^{d} y_i)^p + \sum_{d=k}^{D-1}(1-\sum_{i=1}^{d} y_i)^p > \sum_{d=1}^{D-1}(\sum_{i=1}^{d} y_i)^p \tag{23}$$

After some manipulations (22) and (23) lead to

$$\sum_{d=1}^{k-1}\left[(\sum_{i=1}^{d} y_i)^p - (1-\sum_{i=1}^{d} y_i)^p\right] > 0 \tag{24}$$

$$\sum_{d=k}^{D-1}\left[(1-\sum_{i=1}^{d} y_i)^p - (\sum_{i=1}^{d} y_i)^p]\right] > 0 \tag{25}$$

If relation (24) holds, $(\sum_{i=1}^{d} y_i) > (1-\sum_{i=1}^{d} y_i)$ hence $(\sum_{i=1}^{d} y_i) > 0.5$ for at least one d between 1 and $k-1$. Likewise, from (25), $(\sum_{i=1}^{d} y_i) < 0.5$ for at least one d between k and $D-1$. This is impossible, since $\sum_{i=1}^{d} y_i$ is an increasing function of d and hence (16) holds.

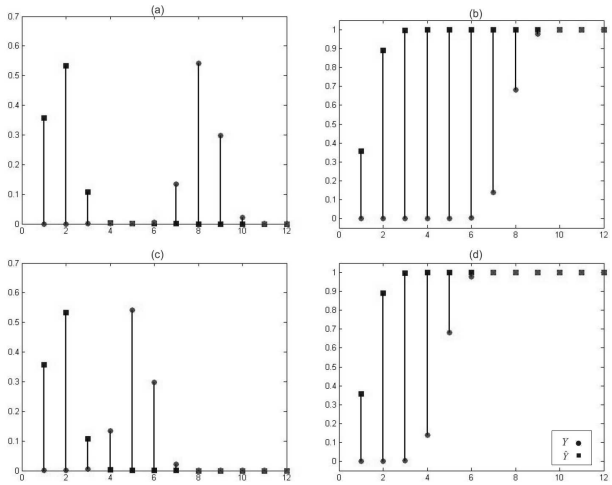

Figure 2: *The effect of using cumulative mass distribution.*

5. RESULTS AND DISCUSSION

In this section, different characteristics of E^{p}_{nod} is discussed and its advantages to conventional performance metrics, namely E_{mzo}, P_{cfci}, E_{ad}, E_{rps}, and $E^{\text{cil}}_{\text{ma}}$ are demonstrated.

5.1. CUMULATIVE PROBABILITY MASS DISTRIBUTION

As it can be interpreted from the relation 13, E^{p}_{nod} calculates the ordinal distance between between $\hat{Y}$ and Y, which is equivalent to Minkowski distance between cumulative probability mass distributions (CMD) of $\hat{Y}$ and Y), hence the order of categories is important. The effect of using CMD is shown in Figure 2 by comparing two cases. Figures 2-a and 2-b show the probability mass distributions (MD) and the CMD of $\hat{Y}$ and Y respectively for case 1. Figures 2-c and 2-d illustrate the MD and the CMD of $\hat{Y}$ and Y respectively for case 2. As it is shown in these figures, $\hat{Y}$ and Y are closer to each other in the second case compared to the first case. While the Minkowski distance between the MD of $\hat{Y}$ and Y does not reflect this fact, the Minkowski distance between CMD of $\hat{Y}$ and Y (ordinal distance of them) shows this closeness effectively.

5.2. ORDER OF CATEGORIES

In example 1, it is shown that P_{cfci} and E_{mzo} are not suitable for measuring the performance of ordinal classifiers, because these methods do not consider the order of categories.

Example 1: For an ordinal three-class classification problem, classifier 1 and classifier 2 result in confusion matrix 1, labeled as CM_1 and CM_2 respectively. In these matrices each column represents the instances in a predicted class and each row shows the instances in an actual class.

$$\text{CM}_1 = \begin{bmatrix} 4 & 1 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 10 \end{bmatrix} \text{CM}_2 = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 5 & 0 \\ 0 & 0 & 10 \end{bmatrix} \tag{26}$$

Table 1: The performance of two classifiers measured by E_{mzo}, E_{ad}, $E^{\text{cil}}_{\text{ma}}$, P_{cfci}, E_{rps}, E^{1}_{nod}, E^{2}_{nod}, and E^{∞}_{nod} in example 1.

Performance Metric	Problem 1	Problem 2
E_{mzo}	0.05	0.05
E_{ad}	0.05	0.1
$E^{\text{cil}}_{\text{ma}}$	0.05	0.1
P_{cfci}	97.5	97.5
E_{rps}	0.025	0.05
E^{1}_{nod}	0.0286	0.0571
E^{2}_{nod}	0.0381	0.0540
E^{∞}_{nod}	0.05	0.05

Table 1 shows the performance of two classifiers measured by E_{mzo}, E_{ad}, $E^{\text{cil}}_{\text{ma}}$, P_{cfci}, E_{rps}, E^{1}_{nod}, E^{2}_{nod}, and E^{∞}_{nod}. As it can be interpreted from this table, E_{mzo}, P_{cfci} and E^{∞}_{nod} fail to reflect the degradation of performance from the classifier 1 to the classifier 2. However, E^{1}_{nod}, E^{2}_{nod}, E_{ad}, E_{rps} and $E^{\text{cil}}_{\text{ma}}$ perfectly show that classifier 1 outperforms classifier 2.

5.3. NUMBER OF CATEGORIES

In Examples 2, it is shown that the number of categories in the classification problem influences the interpretation of E_{ad} and $E^{\text{cil}}_{\text{ma}}$.

Example 2: Consider the following three ordinal and partial ordinal classification problems.

Problem 1: For a test datapoint, the true label and the estimated label are $Y_1 = [1\ 0]$ and $\hat{Y}_1 = [0\ 1]$ respectively.

Problem 2: For a test datapoint, the true label and the estimated label are $Y_1 = [0\ 0\ 0\ 0\ 0.5\ 0.5\ 0\ 0\ 0\ 0]$ and $\hat{Y}_1 = [0\ 0\ 0.5\ 0.5\ 0\ 0\ 0\ 0\ 0\ 0]$ respectively.

Problem 3: In this problem, each two neighboring categories of Y_1 in problem 2 are merged such that the new true and estimated labels are
$Y_1 = [0\ 0\ 1\ 0\ 0]$ and $\hat{Y}_1 = [0\ 1\ 0\ 0\ 0]$ respectively.

Table 2 shows the performance of classifiers in these problems obtained using E_{ad}, $E^{\text{cil}}_{\text{ma}}$, E_{rps}, E^{1}_{nod}, E^{2}_{nod}, and E^{∞}_{nod} in example 2. As it can be interpreted from Table 2, E_{ad}, $E^{\text{cil}}_{\text{ma}}$ and E^{∞}_{nod} treated the classifiers of first and third problems in the same manner. However, the estimated label of the first problem is completely incorrect, while the estimated label in the third problem are near to the true label. Performance metrics E_{rps}, E^{1}_{nod} and E^{2}_{nod} reflect the higher performance of the third classifier compared to the first one.

The second and third problem are naturally similar to each other because the categories in the third problem is obtained by merging the neighboring categories in the second problem. An appealing characteristic of a performance metric is remaining invariant to the number of classes. It can be interpreted from Table 2 that the calculated performance using E_{ad} , E_{rps}, E^{1}_{nod} and E^{2}_{nod} are changed by 200%, 32%, 11% and 16% from problem 3 to problem 2. Therefore, E^{1}_{nod} and E^{2}_{nod} are robust against variability in the number of classes.

5.4. RELATION TO RANKED PROBABILITY SCORE

There is a close relationship between E_{rps} and E^{p}_{nod}, especially for $p = 2$. In both E_{rps} and E^{p}_{nod}, denominators are assumed to be the upper bound of the numerator and are used to keep the range of performance metric between 0 and 1. In E_{rps}, it is assumed that the upper bound of the numerator is $M(D-1)$ (Murphy, 1969; Déqué et al., 1994). However, this is a conservative bound in many situations. In E^{p}_{nod}, this upper bound is explicitly defined by relation (14) and calculated by relation (16). The following examples show that the conservative assumption of E_{rps} results in a misleading or erroneous interpretation of the classifiers performance.

Example 3: Consider the following two cases.

Case 1:

For an ordinal three-class classification problem, a completely useless classifier is applied, which results in CM_3.

$$\text{CM}_3 = \begin{bmatrix} 0 & 0 & 0 \\ 5 & 0 & 5 \\ 0 & 0 & 0 \end{bmatrix} \tag{27}$$

Case 2:

For another ordinal three-class classification problem, consider a classifier with CM_4.

$$\text{CM}_4 = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 10 & 0 & 0 \end{bmatrix} \tag{28}$$

The performance of classifiers in case 1 and 2 calculated by the E_{mzo}, P_{cfci}, E_{ad}, E_{rps}, E^{p}_{nod} and $E^{\text{cil}}_{\text{ma}}$ are listed in Table 3.

As it can be seen from Table 3, the performance of the applied classifier in case 1 measured by E_{rps} is 0.50, while all estimated labels are incorrect and the classifier is totally useless. The outputs of E^{p}_{nod} and P_{cfci} are 1 and 0 respectively, which appropriately reflects that the applied classifier is useless in this case. The table also indicates that E_{rps}, E_{ad} and $E^{\text{cil}}_{\text{ma}}$ result in the same values for both cases, while we know that the applied classifier in the second case is much more effective than the first one. This is appropriately reflected by E^{1}_{nod}, E^{2}_{nod} and E^{∞}_{nod}.

Example 4: This example shows the disadvantage of E_{rps} in measuring the performance of classifiers in $\text{C}^{\text{Pa}}_{\text{O}}$. Consider that in an ordinal-three-class classification problem a

Table 2: The performance of two classifiers measured by E_{mzo}, E_{ad}, $E^{\text{cil}}_{\text{ma}}$, P_{cfci}, E_{rps}, E^{1}_{nod}, E^{2}_{nod}, and E^{∞}_{nod} in example 2.

Performance Metric	Problem 1	Problem 2	Problem 3
E_{ad}	1	2	1
$E^{\text{cil}}_{\text{ma}}$	1	-	1
E_{rps}	1	0.17	0.25
E^{1}_{nod}	1	0.444	0.50
E^{2}_{nod}	1	0.594	0.71
E^{∞}_{nod}	1	1	1

probabilistic classifier is applied. The test set datapoints along with their corresponding classifier outputs are shown in Table 4. Performance metric E_{rps} result suggests that the classifier error is 0.2667, while it can be concluded form Table 4 that the applied classifier is not useful. In this example, E^1_{nod}, E^2_{nod} and E^∞_{nod} results are 1, 0.73 and 0.60 respectively. Obviously, E^p_{nod} better reflects the performance of the applied probabilistic classifier especially for $p = 1$ compared to E_{rps}.

Example 5: In this example, E_{rps} and E^p_{nod} are evaluated in measuring the performance of two classifiers in a real world C_O problem, namely age group classification from speech recordings (Bahari and Van hamme, 2011a). In this experiment, speech signals of 555 speakers from the N-best evaluation corpus (Van Leeuwen et al., 2009) were used. The corpus contains live and broadcast commentaries, news, interviews, and reports broadcast in Belgium. The speakers of this dataset are categorized in three age categories namely, Young $(18-35)$, Middle $(36-45)$ and Senior $(46-81)$. The number of young, middle and senior speakers in this database are 138, 201 and 216 respectively. Among all speakers of the database, 400 are selected for model training and the rest are used for testing. Two approaches are applied for age group recognition. The first method is a random classifier, where $P(\hat{Y} = [1\ 0\ 0]) = P(\hat{Y} = [0\ 1\ 0]) = P(\hat{Y} = [0\ 0\ 1\ ]) = \frac{1}{3}$. The second approach, which is introduced in (Bahari and Van hamme, 2011a), applies well-known speech processing tools and Supervised Non-Negative Matrix Factorization (SNMF) (Bahari and Van hamme, 2012) to recognize the age of speakers. The resulting confusion matrices of both methods can be

$$\text{CM}_{\text{SNMF}} = \begin{bmatrix} 15 & 15 & 9 \\ 18 & 22 & 16 \\ 9 & 11 & 40 \end{bmatrix} \text{CM}_{\text{random}} = \begin{bmatrix} 13 & 13 & 13 \\ 18 & 18 & 19 \\ 20 & 20 & 20 \end{bmatrix} \tag{29}$$

The results of using performance metrics E_{rps} and E^p_{nod} are listed in Table 5.

A subjective study on the obtained results shows that the SNMF based age group recognizer is more effective than a Random classifier. As it can be interpreted from Table 5, this performance drop is better revealed in E^p_{nod} compared to E_{rps}. In this experiment, the error of the random classifier measured by E_{rps} is only 0.44, which is not rational. By contrast, the results of E^1_{nod}, E^2_{nod} and E^∞_{nod} effectively reflect the nature of the applied Random classifier.

Table 3: The performance of two classifiers measured by E_{mzo}, E_{ad}, $E^{\text{cil}}_{\text{ma}}$, P_{cfci}, E_{rps}, E^1_{nod}, E^2_{nod}, and E^∞_{nod} in example 3.

Performance Metric	Case 1	Case 2
E_{mzo}	1	0.5
E_{ad}	1	1
$E^{\text{cil}}_{\text{ma}}$	1	1
P_{cfci}	0	50
E_{rps}	0.50	0.50
E^1_{nod}	1	0.5714
E^2_{nod}	1	0.5395
E^∞_{nod}	1	0.5

Table 4: Test set datapoints and their corresponding classifier outputs in example 4.

	Actual Label(Y)			Classifier Output($\hat{Y}$)		
Datapoint 1	0	1	0	0.3	0	0.7
Datapoint 2	0	1	0	0.6	0	0.4
Datapoint 3	0	1	0	0.5	0	0.5

5.5. PARTIAL-ORDINAL PROBLEMS

Examples 6 and 7 show the advantages of E^{p}_{nod} over P_{cfci}, E_{rps} and E_{ad} in measuring the performance of the classifiers in $\text{C}^{\text{Pa}}_{\text{O}}$, where other conventional approaches are not applicable.

Example 6: In this example, P_{cfci}, E_{ad}, E_{rps}, and E^{p}_{nod} are evaluated in measuring the performance of classifiers in $\text{C}^{\text{Pa}}_{\text{O}}$. Consider an eight-class $\text{C}^{\text{Pa}}_{\text{O}}$. In this problem, the test datapoint label is $Y = [0.1\ 0.1\ 0.1\ 0.1\ 0.1\ 0.1\ 0.2\ 0.2]$. Two classifiers are applied in this problem. Table 6 shows the output of the applied classifiers. The measured performance of these classifiers using P_{cfci}, E_{ad}, E_{rps}, E^{1}_{nod}, E^{2}_{nod} and E^{∞}_{nod} is presented in Table 7. As it can be understood from Table 6, the estimated label of the second classifier is more similar to the true label compared to that of first classifier. However, the output of the P_{cfci} is the same for both of them. This is due to the fact that the order of categories has no effect on the output of P_{cfci}. In this example, E_{ad}, E_{rps}, E^{1}_{nod}, E^{2}_{nod} and E^{∞}_{nod} reflect the performance improvement from the first classifier to the second one.

Example 7: In this example, the behavior of E^{p}_{nod} and E_{rps} in a $\text{C}^{\text{Pa}}_{\text{O}}$ is analyzed. Consider a five-class $\text{C}^{\text{Pa}}_{\text{O}}$. In this problem, a special classifier is applied to recognize the labels of an infinite number of datapoints. The actual label of all datapoints is the same $Y = [0.2\ 0.2\ 0.2\ 0.2\ 0.2]$.
The applied classifier is random and crisp in which $P(\hat{Y} = [1\ 0\ 0\ 0\ 0]) = P(\hat{Y} = [0\ 1\ 0\ 0\ 0]) = P(\hat{Y} = [0\ 0\ 1\ 0\ 0]) = P(\hat{Y} = [0\ 0\ 0\ 1\ 0]) = P(\hat{Y} = [0\ 0\ 0\ 0\ 1]) = 0.2$. The error of the applied classifier expressed by the E_{rps} is 0.20. However, since the classifier is absolutely random, this result is not rational. The measured error using E^{1}_{nod}, E^{2}_{nod} and E^{∞}_{nod} is 0.80, 0.7983 and 0.80 respectively, which perfectly matches the characteristics of this classifier.

6. CONCLUSION

In this paper, the ordinal distance between two arbitrary vectors in Euclidean space has been introduced. Then, Normalized Ordinal Distance (E^{p}_{nod}) as an application-independent performance metric for ordinal, probabilistic-ordinal or partial-ordinal classification problems has been presented. Different advantages of the E^{p}_{nod} over

Table 5: The performance of two classifiers measured by E_{rps}, E^{1}_{nod}, E^{2}_{nod}, and E^{∞}_{nod} in example 5.

Performance Metric	SNMF	Random
E_{rps}	0.30	0.44
E^{1}_{nod}	0.37	0.54
E^{2}_{nod}	0.41	0.60
E^{∞}_{nod}	0.46	0.67

conventional performance metrics such as mean absolute error of consecutive integer labels $E_{\text{ma}}^{\text{cil}}$, mean zero-one error ($E_{\text{mzo}}$), correctly fuzzy classified instances (P_{cfci}), average deviation (E_{ad}), or ranked probability score (E_{rps}) have been shown using a number of numerical examples.

7. ACKNOWLEDGEMENTS

This work is supported by the European Commission as a Marie-Curie Initial Training Networks project (FP7-PEOPLE-ITN-2008), namely Bayesian Biometrics for Forensics (BBfor2), under Grant Agreement number 238803.

The authors also would like to thank Jort F. Gemmeke for his helps to accomplish this work.

References

Bahari, M., Dehak, N., Van hamme, H., B.L., Ali, A., Glass, J., 2014. Non-negative factor analysis of gaussian mixture model weight adaptation for language and dialect recognition. IEEE/ACM Transactions on Audio Speech and Language Processing 22, 1117–1129.

Bahari, M., Van hamme, H., 2011a. Age and gender recognition from speech patterns based on supervised non-negative matrix factorization. 20th annual conference of the international association of forensic phonetics and acoustics , 3–5.

Bahari, M., Van hamme, H., 2012. Speaker age estimation using hidden markov model weight supervectors, in: Proc. 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), pp. 517–521.

Bahari, M.H., Van hamme, H., 2011b. Speaker age estimation and gender detection based on supervised non-negative matrix factorization, in: Proc. IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS), pp. 1–6.

Bougeault, P., 2003. The wgne survey of verification methods for numerical prediction of weather elements and severe weather events. Toulouse: Météo-France .

Cardoso, J., da Costa, J., 2007. Learning to classify ordinal data: the data replication method. Journal of Machine Learning Research 8, 6.

Cheng, J., Wang, Z., Pollastri, G., 2008. A neural network approach to ordinal regression, in: Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, pp. 1279–1284.

Chu, W., Ghahramani, Z., 2004. Gaussian processes for ordinal regression. Journal of Machine Learning Research 6, 1019–1041.

Table 6: The output of applied classifiers in example 6.

	Classifier Output ($\hat{Y}$)							
Classifier 1	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1
Classifier 2	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.1

Table 7: The performance of two classifiers measured by P_{cfci}, E_{ad}, E_{rps}, E^1_{nod}, E^2_{nod} and E^∞_{nod} in example 6.

Performance Metric	Classifier 1	Classifier 2
E_{ad}	1.2	0.2
P_{cfci}	80	80
E_{rps}	0.0314	0.0029
E^1_{nod}	0.2927	0.0488
E^2_{nod}	0.2828	0.0853
E^∞_{nod}	0.2222	0.1111

Chu, W., Keerthi, S., 2007. Support vector ordinal regression. Neural Computation 19, 792–815.

Déqué, M., Royer, J., Stroe, R., France, M., 1994. Formulation of gaussian probability forecasts based on model extended-range integrations. Tellus A 46, 52–65.

Deza, M., Deza, E., 2009. Encyclopedia of distances. Springer.

Erdural, S., 2006. A method for robust design of products or processes with categorical response. METU, Ankara .

Kincaid, D., Cheney, E., 2002. Numerical analysis: mathematics of scientific computing. volume 2. Amer Mathematical Society.

Kohonen, J., Suomela, J., 2005. Lessons learned in the challenge: making predictions and scoring them. Lecture Notes in Artificial Intelligence , 95–116.

Li, M., Han, K.J., Narayanan, S., 2013. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. Computer Speech and Language 27, 151 – 167.

Luxenburger, J., 2008. Modeling and Exploiting User Search Behavior for Information Retrieval. Ph.D. thesis. PhD thesis, Universität des Saarlandes.

Manel, S., Williams, H., Ormerod, S., 2002. Evaluating presence–absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38, 921–931.

McCullagh, P., 1980. Regression models for ordinal data. Journal of the royal statistical society. Series B (Methodological) , 109–142.

Morovic, J., Shaw, J., Sun, P., 2002. A fast, non-iterative and exact histogram matching algorithm. Pattern Recognition Letters 23, 127–135.

Mouton, A., De Baets, B., Van Broekhoven, E., Goethals, P., 2009. Prevalence-adjusted optimisation of fuzzy models for species distribution. Ecological Modelling 220, 1776–1786.

Murphy, A., 1969. On the ranked probability score. J. Applied Meteorology 8, 988–989.

Shevade, S., Chu, W., 2006. Minimum enclosing spheres formulations for support vector ordinal regression, in: Data Mining, 2006. ICDM'06. Sixth International Conference on, IEEE. pp. 1054–1058.

Toda, M., 1963. Measurement of subjective probability distributions. Technical Report. DTIC Document.

Van Broekhoven, E., Adriaenssens, V., De Baets, B., 2007. Interpretability-preserving genetic optimization of linguistic terms in fuzzy models for fuzzy ordered classification: An ecological case study. International Journal of Approximate Reasoning 44, 65–90.

Van Leeuwen, D.A., Kessens, J., Sanders, E., Van Den Heuvel, H., 2009. Results of the n-best 2008 dutch speech recognition evaluation. NOVA 6, 11–5.

Verwaeren, J., Waegeman, W., De Baets, B., 2012. Learning partial ordinal class memberships with kernel-based proportional odds models. Computational Statistics and Data Analysis 56, 928–942.